MODELING SPATIAL DATA: CHALLENGES AND APPLICATIONS TO SOLAR IMAGERY

Michael Turmon
18 January 2002

A. Scientific inference

B. Modeling observations

C. Spatial models

D. Hierarchical and spatiotemporal models

E. Outlook

Work with Judit Pap, U. Maryland, and Kacie Shelton, JPL

turmon@aig.jpl.nasa.gov
http://www-aig.jpl.nasa.gov/home/turmon/

ML FOR SCIENTIFIC INFERENCE

Machine learning methods always give:

Automation: Mechanized process reduces labor and time needed
- Cope with increasing data volume (instruments, simulations)
- Important for data centers: operations often underfunded

Repeatability: Well-defined algorithm produces results
- Uniformity over time key for long-term studies
- Allows uniformity among distributed investigators
- Crucial for highly charged subjects like climate change

Sometimes one obtains these as well:

Objectivity: Problem-sensitive decision among many conclusions
- E.g., model order, number of clusters, which features to use
- Often only possible in a limited context or domain

Consensus: Ubiquitous algorithms factor out disagreements
- Go beyond ad hoc gadgets to general, cross-domain solutions
- Exchange models and algorithms as well as data

Composability: Can analyze machine-generated interpretations
- Building a data pipeline, meta-analysis, federated databases

Performance gains are important:

Quality: Quantitative, optimal inference gives better results
- Many schemes (implicitly) optimize over interpretations
- Gauss obtained the orbit of Ceres by least squares in 1801

Comprehensiveness: Ability to examine more information
- Integrate more data within a given interpretation
- Achieve total spatial/temporal coverage

APPROACHES TO INFERENCE

Generative methods model the data generation process and thus capture the full statistics of the data. Invert models statistically (e.g., by ML or Bayes) to find parameters of interest.

Linear regression by least squares

Remote sensing models

Discriminative methods use a complex-enough system to mimic expert behavior, often as captured in training data.

Linear discriminant analysis (LDA)

Neural network digit recognition

Algorithmic methods follow a reasonable procedure to deduce information from data.

Nearest neighbors

Boosting

Theoretical support for discriminative methods might come by ensuring the system has enough representational capacity for its task: Barron's results on neural network approximation error.

Theoretical support for algorithmic methods sometimes develops after the method meets with success: Cover and Hart's 1967 results for nearest neighbors, and recent work on the relation of boosting to classifier margin.

GROUND TRUTH — MODEL VALIDITY

Questions brought to fore by scientific problems

Physical questions that seem decidable in principle...

...but whose very intractability motivates inference techniques!

Models for observables

Observables are directly sensed, allowing direct model checks

Can falsify (Popper 1958), but never fully verify

Computing $P(\text{data} \mid \text{model})$ falsifies some models or model classes

E.g., image modeled as three classes, each of which is normal, is falsified if pooled pixels are not a normal three-mixture

Information on hidden variables

This 'ground truth' is difficult to come by

- Scientists typically cannot identify objects reliably
  Problems become very evident at single-pixel scale
  The most informative test cases are also most uncertain
- Further: Lack of physical understanding of problem means even experts may be surprised at what is really there.

Conceptual inadequacies in models

Methods are often not suitably invariant to resolution

Classes in image segmentation are often not mutually exclusive

Spatial independence is often assumed at some point

Need spatial/temporal stationarity which rarely exists

Bayesian 'dogma of precision': every state can be assigned a probability; every outcome can be assigned a cost (Walley 1991)

IMAGE LABELING

Solar imagery

Reliably identify structures in the photosphere
- Sunspots: Depressed intensity and high magnetic flux
- Faculae: Regions of enhanced intensity and moderate flux
- Quiet sun: everything else

Relate these structures to irradiance changes (weather/climate)
Also: space weather (identify large δ-spots which cause flares)

Mars Geology

Identify soil structure (dust, sand, pebbles)
Detect rocks on soil background
Classify rock types (sedimentary/igneous, weathering, impact)

Methods

Automatic, objective classification using statistical model

Model quantifies the uncertain relation of observables to classes

Model uses spatial information to choose labels

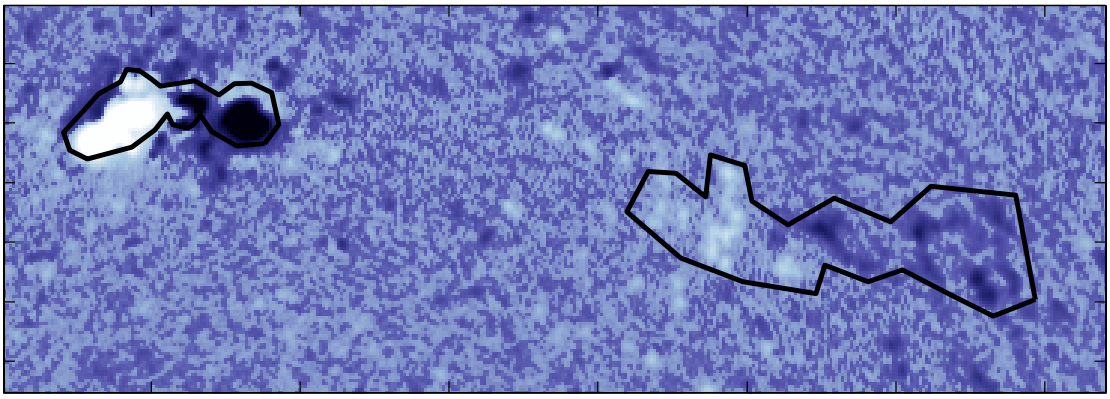
Falsifiable models can be checked against
- the data they claim to model

General method that extends unchanged to other settings, e.g.
- more observables
- different number of features
- explicit accounting for miscalibration; outliers
- inclusion of physical knowledge (like sensor noise)

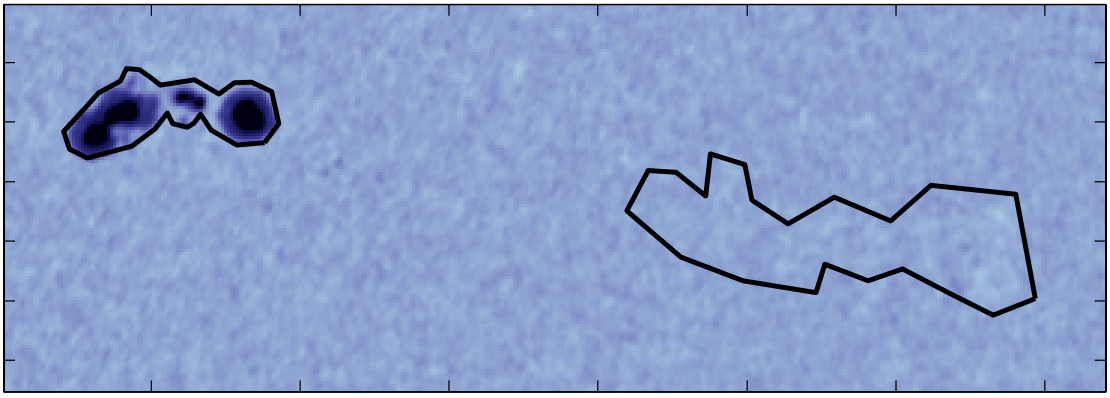
EXAMPLE SOLAR DATA

Irregularly-sampled time series of (full-disk) images
Analyzed May 1996 – Sep 2000; 60 GB across 25 000 images
Below: SoHO/MDI, 17:58 UTC on 7 September 1997

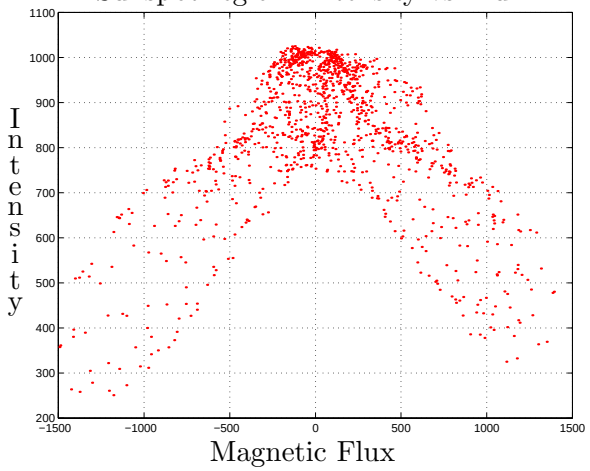
Preprocessed Magnetogram: Detail



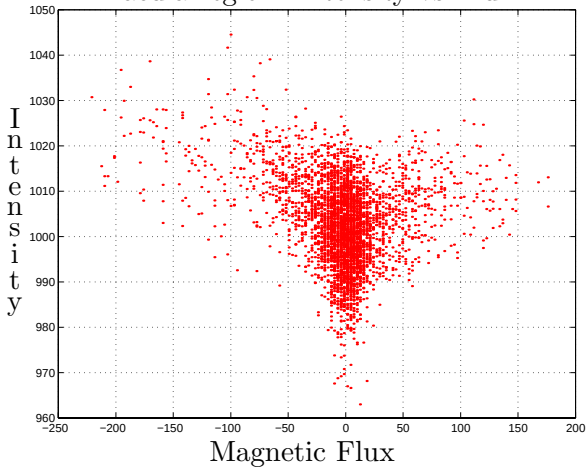
Preprocessed Photogram: Detail



Sunspot region: Intensity vs. flux

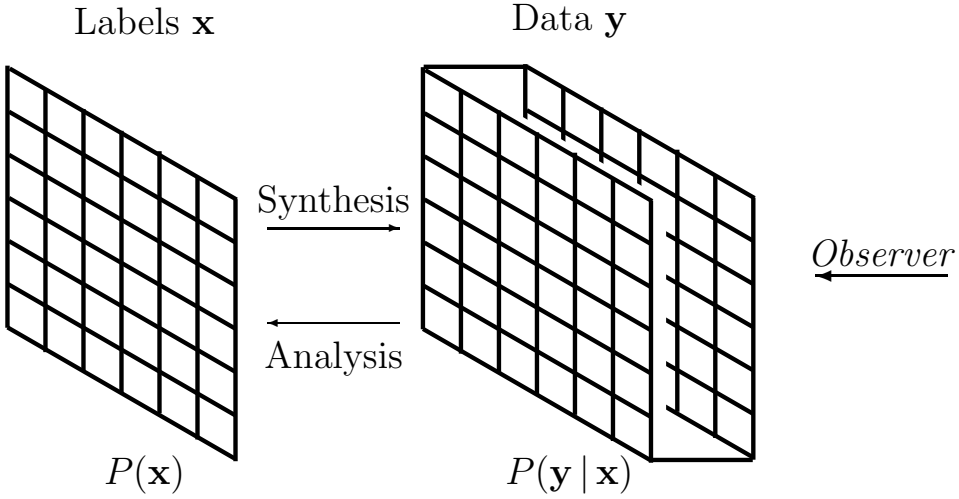


Facula region: Intensity vs. flux



PROBABILISTIC IMAGE MODELS

Quantitatively describe the uncertain relation between observables and labels in a general probabilistic framework



At each spatial position, one of K physical processes is dominant.

Observables arise depending on the dominant physical process.

Generation of observables may be viewed as adding uncertainty (noise) to the underlying dominant process.

Goal of analysis is to invert this noisy mapping.

Variables of the Model

Index set $\mathcal{N}$ of spatial coordinates $s = (i, j)$

Unobservable labels $\mathbf{x} = [x_s]_{s \in \mathcal{N}}$ & observables $\mathbf{y} = [\vec{y}_s]_{s \in \mathcal{N}}$

x_s: small integer $1 \ldots K$ (e.g., ACR/Fac/QS)

$\vec{y}_s$: real vector (e.g., the pair (magnetic field, light intensity))

Statistical model given by two distributions $P(\mathbf{x})$ and $P(\mathbf{y} \mid \mathbf{x})$

MODEL THE OBSERVATIONS

Linking to Observables with $P(\mathbf{y} \mid \mathbf{x})$

Make the link via scientist-labeled images and distribution-fitting

Alternatively, can infer automatically from data via clustering

Obtain K distributions, one for each feature class

As strawman, put forward per-class normal distributions

$$P(\vec{y}_s \mid x_s = k) \sim \text{Normal}(\vec{\mu}_k, \Sigma_k)$$

with $d \times 1$ class means and $d \times d$ covariance matrices.

(QS class, $k = 1$: fits the SoHO/MDI data reasonably well using $\vec{\mu}_1 = [\,0 \;\; 1\,]$ and $\Sigma_1 = (0.01)^2 I$.)

For MDI, the normal distribution is inadequate for all classes:

strongly multimodal
cannot even transform to normality (e.g., with |flux|)
quiet class,e.g., contains superpositions of effects
(supergranulation is discernable in scatter plots)
$\implies$ it fails standard statistical tests.
...normal model is thus *falsified*.

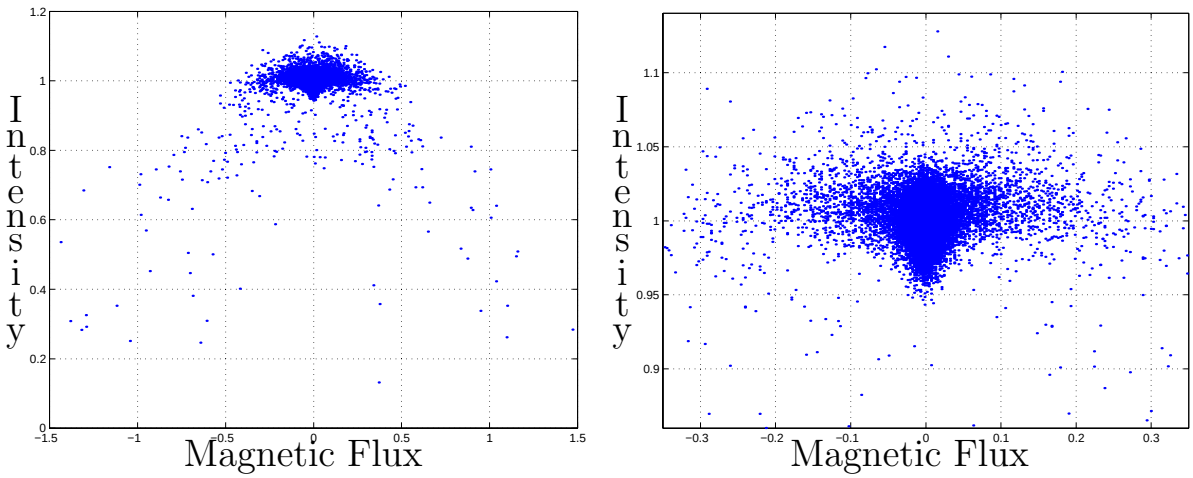
We must introduce more realistic data models $P(\vec{y} \mid x)$

MIXTURE DENSITY MODELS

$$p(\vec{y};\theta) = \sum_{g=1}^{G} \alpha_g N(\vec{y};\ \vec{\mu}_g, \Sigma_g)$$

$$\theta = \{(\alpha_1,\ \vec{\mu}_1,\ \Sigma_1) \cdots (\alpha_G,\ \vec{\mu}_G,\ \Sigma_G)\}$$

Accounts for multimodality and superpositions of effects
A very general family: take G large.



Goal: From data $Y = [\vec{y}^1 \cdots \vec{y}^n]$, find a density model $p(\vec{y};\hat{\theta})$
Method: Determine parameters by maximum-likelihood using Y:

$$\hat{\theta} = \arg\max_{\theta} \log P(Y;\ \theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log p(\vec{y}^i;\ \theta)$$

Performed via EM algorithm: done once and the model is fixed

Supervised mode: scientists find regions of each class $x_s = k$; estimate θ_k independently for $1 \le k \le K$
Unsupervised: Provide pooled data, then EM divides $\vec{y}$ into components — assign G bumps to K classes after the fact.
Mixed mode: Some class-assignments are supplied to EM, others are determined in unsupervised mode.

MODEL VALIDATION

Overfitting

Find $\hat{\theta}$ from Y, varying number of bumps $G = 1, 2, ...$

$$\hat{\theta}_G = \arg\max_{\theta} \log P_G(Y;\, \theta)$$

As G increases, "better" fits $\hat{\theta}_G$ to Y are obtained

Overfitting phemomenon: too many parameters to fit reliably

Controlling model complexity with cross-validation

Cross-validated likelihood (Smyth 1999)

Solution: evaluate models on a separate validation data set

Hold aside test data $Z = [\vec{y}^1 \cdots \vec{y}^m]$ disjoint from Y

Train $\hat{\theta}_G$ from Y with maximum-likelihood as indicated

Test $\hat{\theta}_G$ on separate data Z

Contrast *test likelihood* $P_G(Z; \hat{\theta}_G)$ with $P_G(Y; \hat{\theta}_G)$: the former is an unbiased estimate of fit of $\hat{\theta}_G$ to true distribution

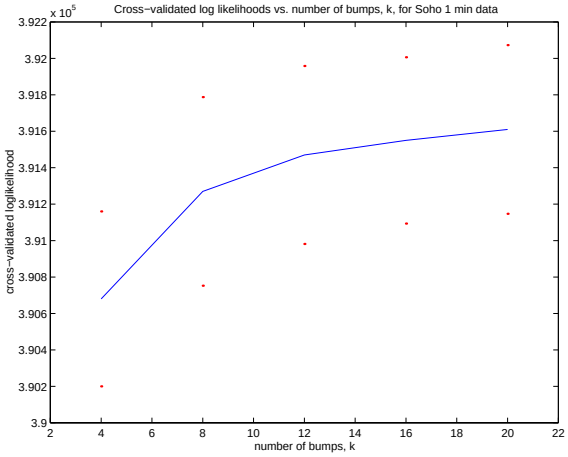
Next, generate more training/test (Y/Z) splits to get more estimates of goodness-of-fit

Average of these goodness-of-fit indicators shows what model complexity the data can support

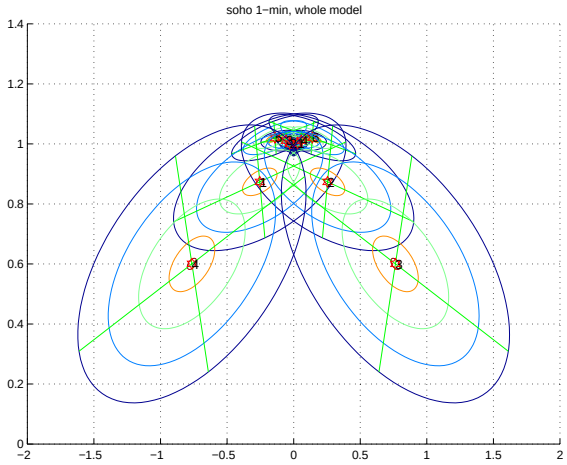
Also, can serve to test robustness of model to changes in data to which it should be largely invariant

MODELS USED: SOHO/MDI

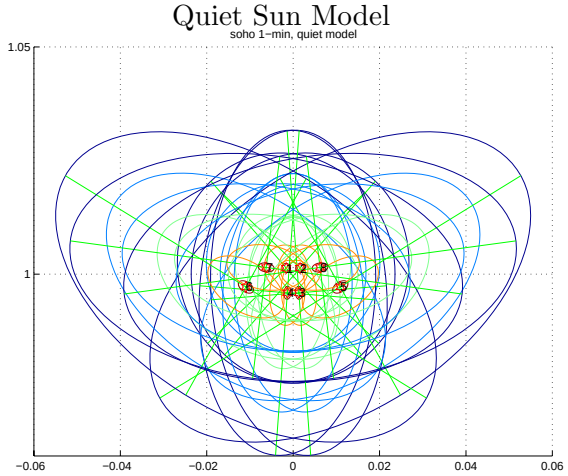
Model Fit, varying Complexity



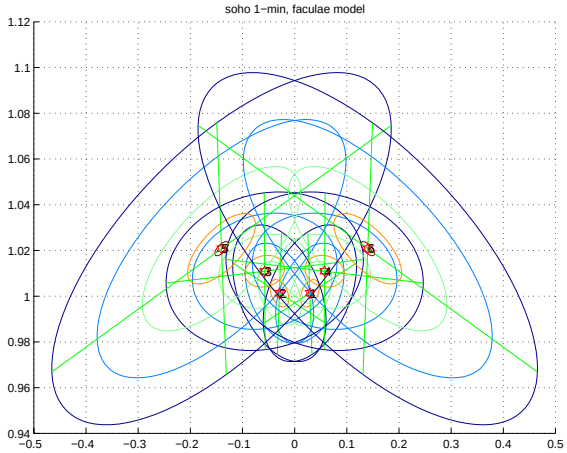
Entire Model



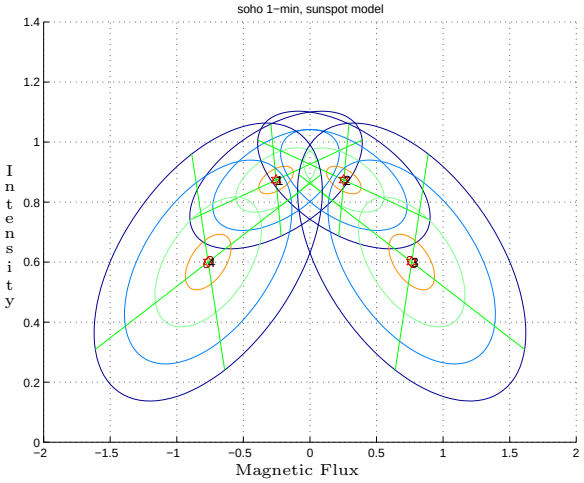
Quiet Sun Model



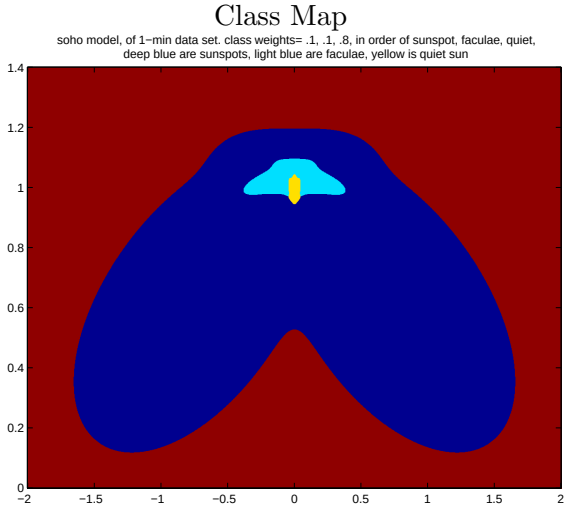
Facula Model



Spot Model

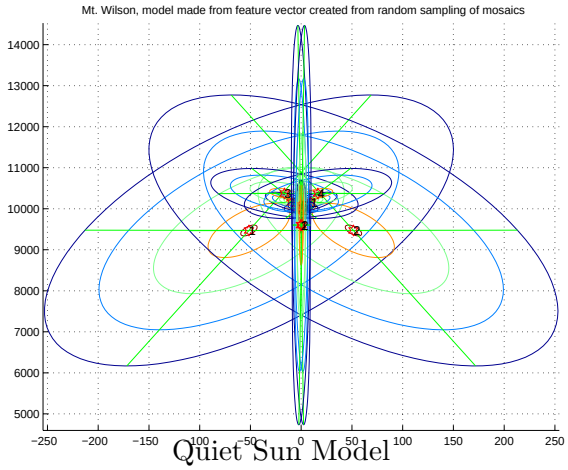


Class Map

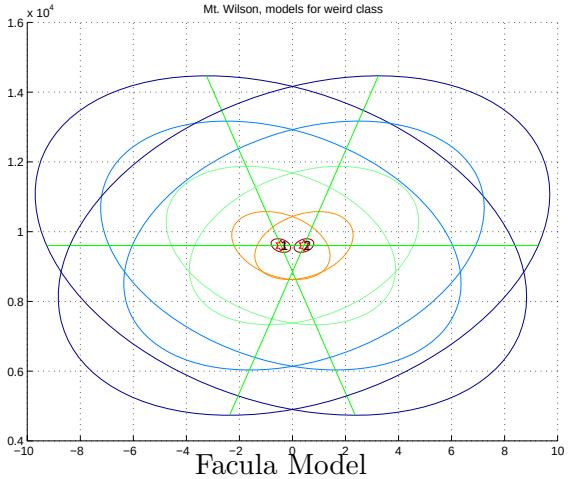


MODELS USED: MT. WILSON

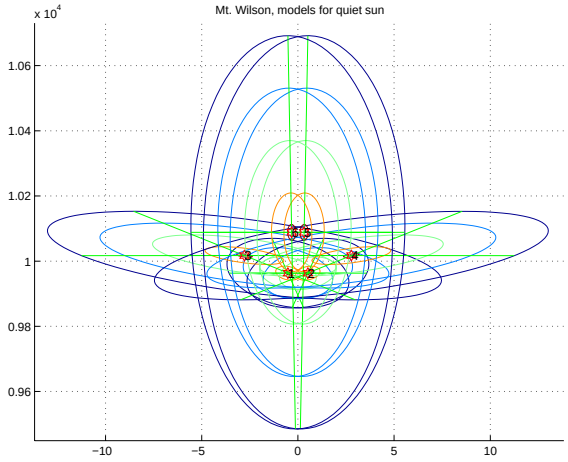
Entire Model



Miscalibration Model

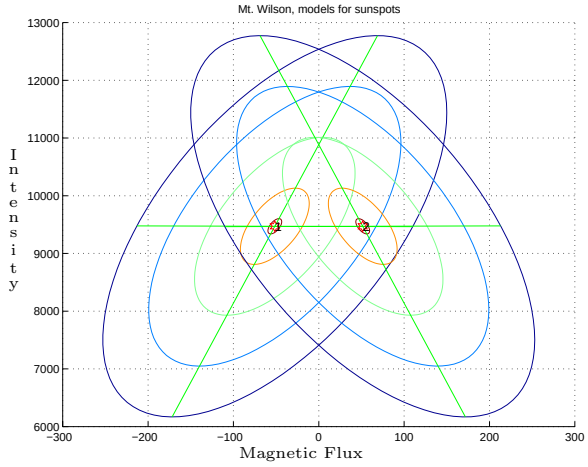


Quiet Sun Model

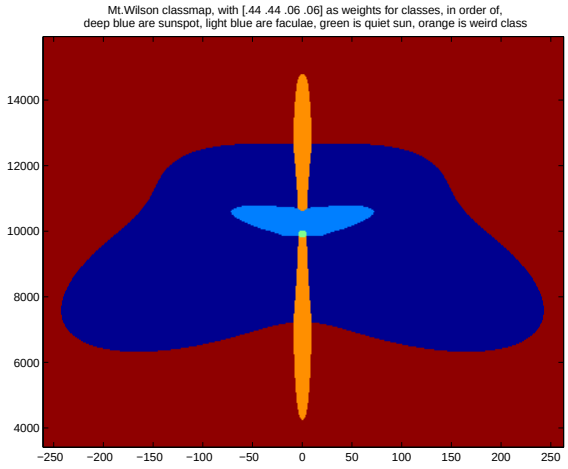


Facula Model

Spot Model



Class Map



MODELING SPATIAL VARIATIONS

Quantifying Spatial Smoothness with $P(\mathbf{x})$

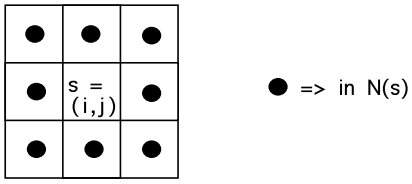
Typically $\beta \geq 0$ controls smoothness in the prior

$$P(\mathbf{x}) = \frac{1}{Z} \exp\Big(-\beta \sum_{s \sim s'} 1(x_s \neq x_{s'})\Big)$$

where $s \sim s'$ means: site s close to site s', e.g. one pixel away

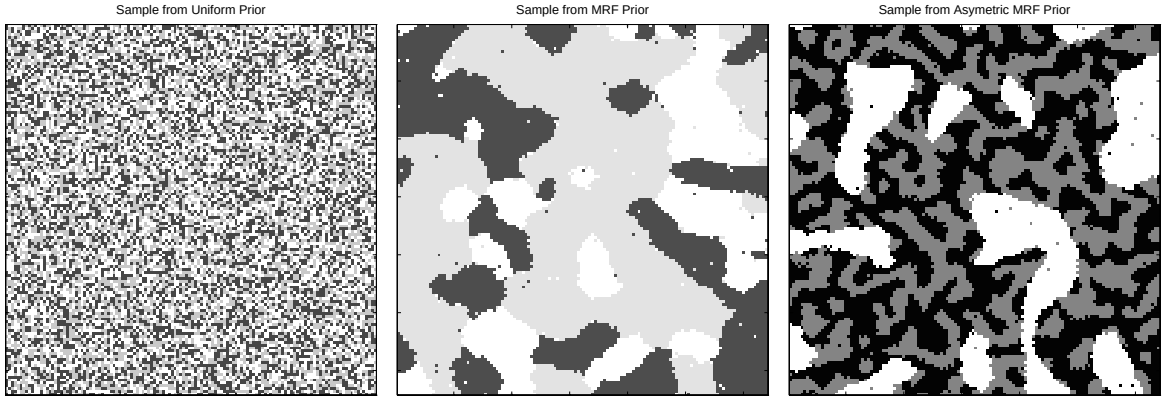
Penalty of β per disagreement of nearby pixels to enforce spatial coherence of labelings

Key property of locality: $P(x_s = x \,|\, x_{(s)}) = P(x_s = x \,|\, x_{\mathcal{N}(s)})$



At $\beta = 0$, penalty and spatial constraint vanish

Sample realizations from $P(\mathbf{x})$



ASIDE: CONTINUITY AND EDGES

Such Markov random field models allow edges in modeled images

Change in discrete hidden variable forces
significant change in real-valued observable

Jumps undesirable in many image *restoration* contexts

Motivates conditional autoregressive (CAR) model

$$P(x_s = x \,|\, x_{(s)}) = P(x_s = x \,|\, x_{\mathcal{N}(s)}) = N(Ax_{\mathcal{N}(s)}, \Sigma)$$

but with conditionally normal distribution

(Autoregression: predict x_s in terms of "itself" $x_{\mathcal{N}(s)}$)

Joint distribution of CAR model is normal, easing computation

Natural parallel with familiar one-dimensional models

	Continuous	Discrete
Time Series	Autoregressive (AR) or Kalman models	Hidden Markov models (HMM)
Imagery	CAR models	Markov random fields (MRF)

MRF computations are the hardest: our best tools do not apply

Non-gaussian, so no reduction to clever matrix manipulations
Bayes net of many short cycles, junction tree algs liable to fail

But: sampling, Metropolis-Hastings, and MCMC methods developed for MRFs enable very complex models

SIMULATING MRFS

Distribution $P(\mathbf{x}) = Z^{-1} \exp\Big(-\beta \sum_{s\sim s'} 1(x_s \neq x_{s'})\Big)$

No direct simulation: no Z, and state space of $\mathbf{x}$ huge!

Randomized algorithm: Gibbs sampler

Basis: craft a MC having P as its stationary distribution

Adaptation of stat-mech methods (c.f. Metropolis *et al.* 1953) for simulating the state of interacting systems

Iterative algorithm: starts at some labeling and refines it pixel-by-pixel over many image sweeps

Method:

- Choose an initial $\hat{\mathbf{x}}$
- Scan pels in raster fashion.
  - At pel s, find $P(\hat{x}_s = x \,|\, \hat{x}_{(s)})$, $1 \leq x \leq K$.
  - Choose new $\hat{x}_s$ by drawing from this distribution [*]
- Repeat scanning

Result: As scans go to infinity, $\hat{\mathbf{x}} \Rightarrow P(\cdot)$.
That is, iterate enough and the labeling is a draw from $P(\mathbf{x})$

Remarks

Flip of one label can eventually influence all labels

This method, and similar Metropolis-Hastings methods, are the basis for updating more complex spatial models

INFERRING THE LABELING

Invert the noisy data via *maximum a posteriori* (MAP) rule

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} P(\mathbf{x}|\mathbf{y})$$

Bayes formula shows $P(\mathbf{x}|\mathbf{y}) \propto P(\mathbf{y}|\mathbf{x})P(\mathbf{x})$

For normal $P(y_s \,|\, x_s)$, algebra reveals the objective function

$$\log P(\mathbf{x}|\mathbf{y}) = -\frac{1}{2\sigma^2}\sum_{s\in\mathcal{N}} \left\|\vec{y}_s - \vec{\mu}_{x_s}\right\|^2 - \beta \sum_{s\sim s'} 1(x_s \neq x_{s'})$$

Interpretation

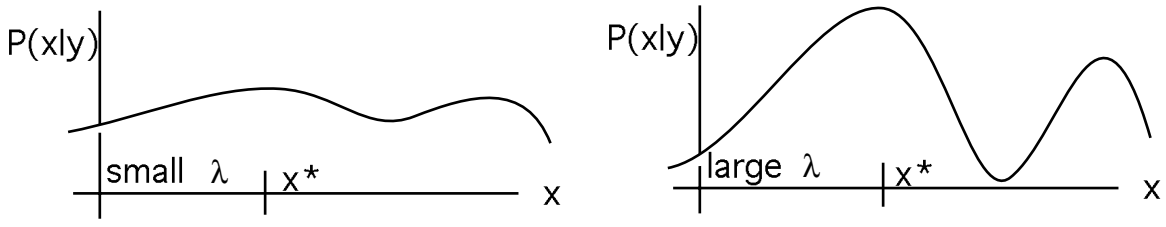
First term: fidelity to data (observation close to its mean)
Second term: image smoothness (this couples the pixel labels)

Maximizing $P(\mathbf{x} \,|\, \mathbf{y})$

- Use Gibbs sampler to *draw* from the distribution $P(\mathbf{x} \,|\, \mathbf{y})$
- To *maximize* $P(\mathbf{x} \,|\, \mathbf{y})$, nest G.S. within simulated annealing
  That is, pick large λ and draw via G.S. from

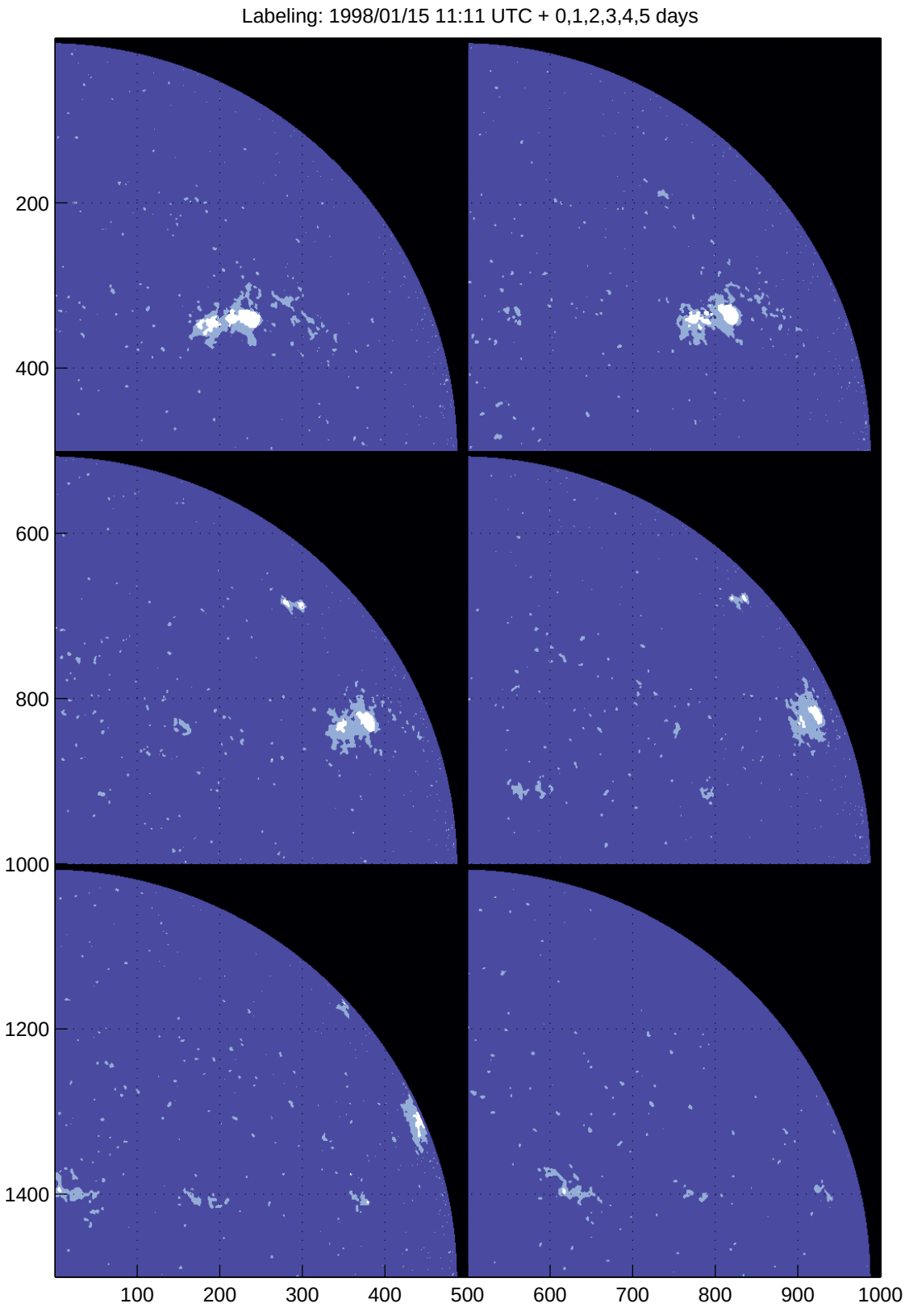
  $$P_\lambda(\mathbf{x} \,|\, \mathbf{y}) := (1/Z_\lambda)\, P(\mathbf{x} \,|\, \mathbf{y})^\lambda$$

  (Effectively scale entire log-posterior, above, by λ)
- Simulated annealing: raise λ as Gibbs sampler iterates
  If λ up slowly enough, mode is reached



- Takes about 3 min/image on Sun workstation (360MHz).

SOHO/MDI LABELINGS



CHALLENGES IN LABELING

Belief propagation

MRF in Bayes network notation:

Belief propagation algorithm finds MAP estimate in acyclic networks — MRF clearly has cycles!

Y. Weiss (2000) found conditions for belief propagation to produce correct answers even for graphs with cycles; these include certain Gaussian models but not discrete MRFs.
Interestingly, Fridman and Mumford used belief propagation successfully for inference in some tightly coupled MRF graphs.

Relaxation labeling methods

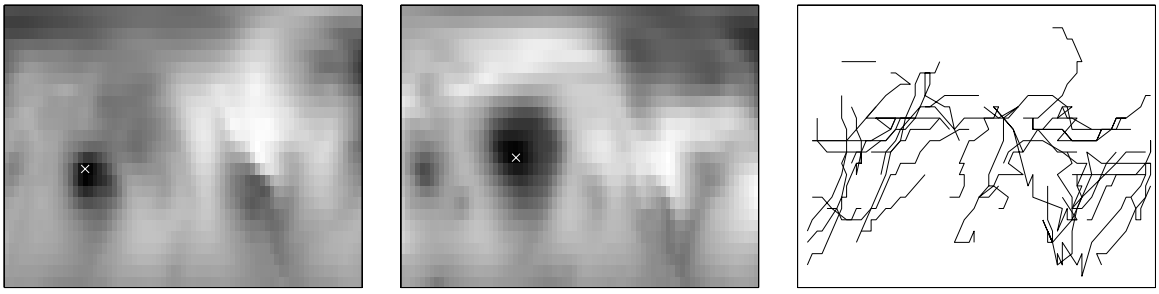
Gibbs sampling algorithms flip pixel labels one at a time.
Can relax the labeling problem into a continuous-valued linear program, and use max-flow/min-cut methods to identify connected groups of labels to flip all at once.

On the one hand, impractical poly-time algorithms come within small factors of the global optimum (Kleinberg and Tardos 1999).
On the other hand, heuristic cut-set identification algorithms are practical for medium-sized problems and have significantly outperformed Gibbs sampling (Boykov, Veksler, Zabin 1998; Ishikawa 2000).

SPATIO-TEMPORAL INFERENCE

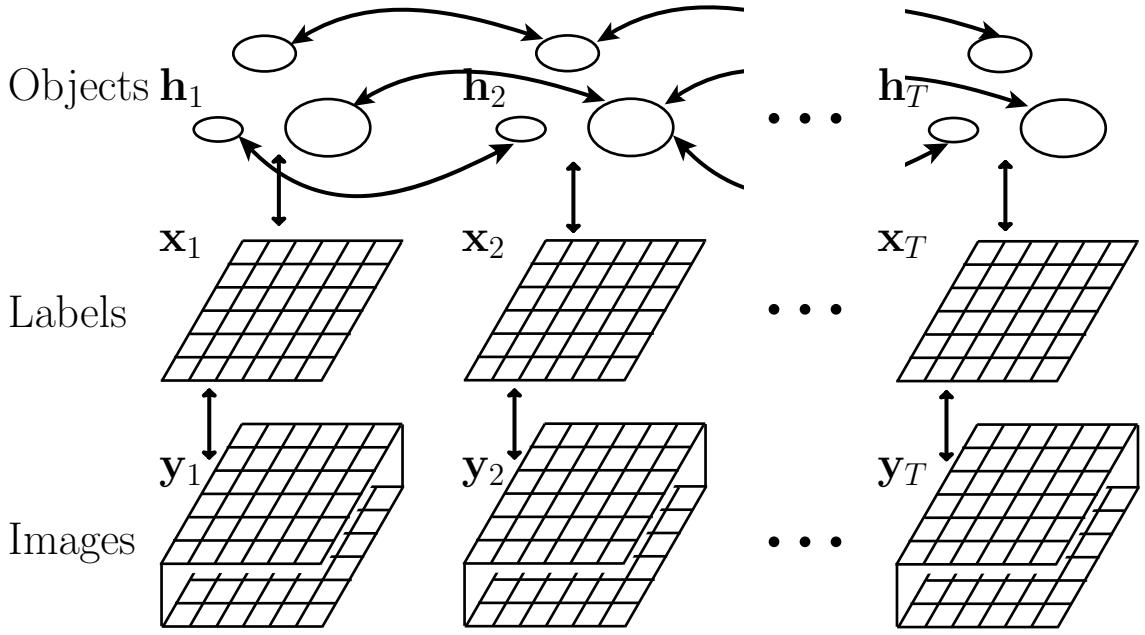
Object trajectories

Sea-level pressure over the Pacific ($\delta t = 48$ hrs.)
Cyclone center shown by white cross
Right: trajectories from a series of (quantized) observations
Data from P. Smyth, UC Irvine



Other examples: sunspot motion, microblock motion from GPS

Objects through time



HIERARCHICAL SPATIAL MODELS

Better Representations

Represent an object via a *compactly-described* membership function h_s indicating subjective belief site s is active region

— Larger-scale representation of an object

— Provides interpretability

Several Simple Mechanisms

Outlines: Grenander et al., 1991
Polygons: Green 1996
Continuum triangulations: Nicholls 1997, 1998
Delaunay triangulations: Turmon 1998

Binds nearby on-object regions into one object

Two fundamental quantities:

Indicator function h_s, $s \in \mathcal{N}$

$h(s) = 1$ means on-object, $h(s) = 0$ if not

Parameterized by tie points in $\mathcal{N}$.

Function complexity $\kappa(h) \geq 0$

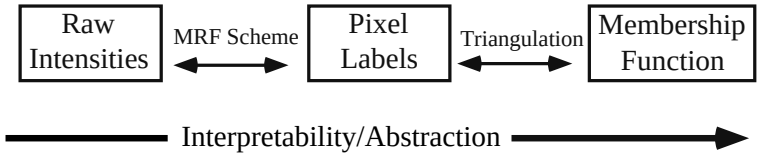
e.g., the number of tie points, or

intensity of point process generating tie points

LINK TO OBSERVATIONS

Establish Markov dependence between hierarchical model layers

$$P(h, \mathbf{x}, \mathbf{y}) = P(h)P(\mathbf{x} \,|\, h)P(\mathbf{y} \,|\, \mathbf{x})$$



Probabilistic Model

Penalize complexity by setting

$$P(h) = Z^{-1} \exp\big[-\gamma\, \kappa(h)\big]$$

This choice gives an additive penalty to disjoint objects

Intermediate layer uses h_s to bias the event $\{x_s = \texttt{Object}\}$:

$$-\log P(\mathbf{x} \,|\, h) = \beta \sum_{s \sim s'} 1(x_s \neq x_{s'}) + \alpha \sum_{s \in \mathcal{N}} |1(x_s = \texttt{Object}) - h(s)|$$

The data distribution $P(\mathbf{y} \,|\, \mathbf{x})$ is as before.

- One can do inference by maximizing the posterior

$$P(h, \mathbf{x} \,|\, \mathbf{y}) = P(h, \mathbf{x}, \mathbf{y})/P(y) \propto P(h, \mathbf{x}, \mathbf{y})$$

or minimizing its negative logarithm

$$\gamma\, \kappa(h) + \alpha \sum_{s \in \mathcal{N}} |1(x_s = \texttt{Object}) - h(s)| + \beta \sum_{s \sim s'} 1(x_s \neq x_{s'}) + \frac{1}{2\sigma^2} \sum_{s \in \mathcal{N}} (y_s - \mu_{x_s})^2$$

INFERRING COMPLEX MODELS

We describe inferring shape models for fixed labeling

To speed convergence, replace $1(x_s = \texttt{Object})$ above with its probability given the data
(Fully analogous to ICE algorithm of Art Owen)

Now the objective simplifies to

$$\gamma\,\kappa(h) + \alpha \sum_{s \in \mathcal{N}} \Big| P(x_s = \texttt{Object} \,|\, \mathbf{y}) - h(s) \Big|$$

Metropolis-Hastings sampler

Inference means choosing tie-point positions

Construct a Markov chain on the state space of tie points

$$\mathcal{V} = \bigcup_k \mathcal{V}_k = \bigcup_k (\mathcal{N} \times \mathcal{N})^k$$

that has limit distribution

$$\pi(h) = P(h \,|\, \mathbf{x}, \mathbf{y})$$

(Maximize $P(h \,|\, \mathbf{x}, \mathbf{y})$ with same annealing setup as earlier)

Metropolis-Hastings proposes state changes and probabilistically accepts them to achieve the desired limit distribution

The operator set consists of
- tie-point move (M),
- tie-point raise/lower (R),
- tie-point add (A_k) or kill (A'_k)

DESIGNING THE MARKOV CHAIN

The operator set consists of

vertex move (M),
vertex raise/lower (R),
vertex add (A_k) or kill (A'_k),
gradient move (G)

Operators are chosen equiprobably. Each one...

At state h, proposes a move to h' with prob. $q(h, h')$
Accepts the proposed state with probability $\alpha(h, h')$

Reaching Equilibrium

Presence of add/kill operators ensures irreducibility of MC.
Positive probability of rejecting the proposal ensures aperiodicity.

Detailed balance is thus sufficient to obtain equilibrium at π
(variable-dimension state space makes this nontrivial)

We pick $\alpha(h, h')$ between 0 and 1 such that

$$\frac{\alpha(h, h')}{\alpha(h', h)} = \frac{\pi(h')q(h', h)}{\pi(h)q(h, h')}$$

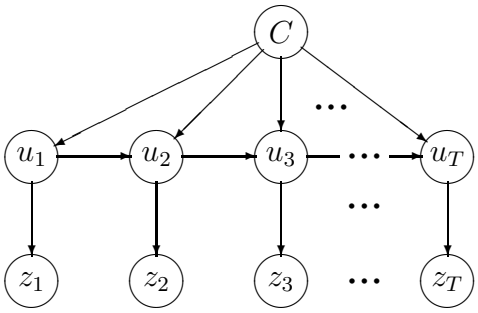
The local changes introduced by the operators make $\pi(h')/\pi(h)$
easy to compute: only changed triangles affect it

MODELING THE TEMPORAL PART

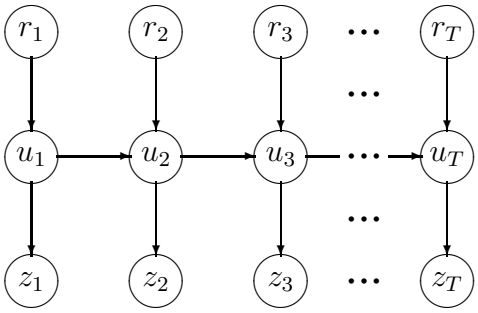
State-based motion models
Include influence of exogenous inputs and observable covariates
Discover motion clusters by uncovering a hidden class C

Examples

Generalizations of the Kalman filter as Bayes nets with state u_t



mixed dynamical model



model with exogenous inputs r_t

Build temporal models atop de-coupled spatial models

Implications

Two domains of divide and conquer
Easy cases: dominant locality in space (sunspots) or time (GPS)
...allows decoupled solutions
Coping with both simultaneously is harder, even beyond
current limits of practical optimization technology

Problems...
estimate model parameters automatically
learn the model structure automatically

CHALLENGES IN OBJECT MODELING

Markov chain Monto Carlo

We have available various sampling techniques, including Metropolis-Hastings, Gibbs sampling, diffusion processes (e.g., Miller, Grenander), and variable-dimension sampling (e.g., P. Green).
It is relatively easy to design a sampler, but hard to show it has converged to the stationary distribution:

subtle identifiability problems
multimodality
flat spots in posterior

Perfect sampling

Can sometimes draw exact samples from Markov chains in which the stationary distribution is reached only asymptotically (Propp and Wilson, 1996).

Such samples drawn from Potts MRF priors have revealed that typical Gibbs sampling MRF simulations are over-smoothed!

Perfect sampling may contribute to better convergence of MCMC samplers (e.g., Green and Murdoch 1999).

CONCLUSIONS

Machine procedures offer many benefits to scientific inference

Persistent issues:

- Building tractable models of observational reality
- Obtaining accurate training data
- Designing and executing clear falsification experiments

Labeling Images

Use of statistical models allows falsification experiments,
easy extension to wider class of problems
Spatially, temporally uniform data is key to accurate labelings

Complex models

Useable temporal and spatial statistical models do exist
...but the best ML perspectives often absent from this work
agnostic models, robust algorithms, cross-validation,
automation
Cooperating space/time models, linked spatiotemporal models

Futures

Languages to express statistical models on structured domains
Model selection in complex, flexible model space